Ⅲ　回帰分析とその統計的性質

母集団：$Y_i=\alpha+\beta X_i+\varepsilon_i$ ―無作為抽出→ 標本(n=100)：$Y_i=a+bX_i+e_i$、e_i(最小２乗残差)

Y_i:成績, X_i:勉強時間, ε_i:誤差　　　何通りもの標本抽出が可能､(a,b)は確率分布する

$$b=\sum w_i Y_i=\beta+\sum w_i\varepsilon_i,\ w_i=\frac{X_i-\bar{X}}{\sum(X_i-\bar{X})^2}\quad (\sum w_i=0, \sum w_i(X_i-\bar{X})=\sum w_i X_i=1)$$

$$a=(\alpha+\beta\bar{X}+\bar{\varepsilon})-b\bar{X}=\alpha-(b-\beta)\bar{X}+\bar{\varepsilon}=\alpha+\sum(\tfrac{1}{n}-w_i\bar{X})\varepsilon_t$$

0. 期待値、分散演算子の計算公式

(1)期待値　$E(X)=\Sigma X_i f(X_i)$ or $E(X)=\int Xf(X)dX$

$E(bx)=bE(x)$, $E(a+x)=a+E(x)$, $E(a+bx)=a+bE(x)$

(2)分散　$V(X)=E\{(X-E(X))^2\}=\Sigma(X_i-E(X_i))^2 f(X_i)$ or $E(X)=\int (X-E(X)^2 f(X)dX$

$V(bx)=b^2V(x)$, $V(a+x)=V(x)$, $V(a+bx)=b^2V(x)$

(3)共分散　$Cov(X,Y)=E\{(X-E(X))(Y-E(Y))\}$

$Cov(X,X)=V(X)$, $Cov(X,Y)=Cov(Y,X)$, $Cov(a+X,Y)=Cov(X,Y)$, $Cov(bX,Y)=bCov(X,Y)$

1. 誤差項 ε 、説明変数 X の標準的仮定（p-25-27）

(U1) ε_iの期待値はゼロ:$E(\varepsilon_i)=0$

(U2) ε_iの分散均一性:$Var(\varepsilon_i)=E\{(\varepsilon_i-E(\varepsilon_i))^2\}=E(\varepsilon_i^2)=\sigma^2$,　σ:標準誤差(standard error)

(U3) ε_iは互いに無相関:$Cov(\varepsilon_i,\varepsilon_j)=E\{(\varepsilon_i-E(\varepsilon_i))(\varepsilon_j-E(\varepsilon_j))\}=E(\varepsilon_i\varepsilon_j)=0$

(U4) ε_i正規分布に従う(中心極限定理):U1-4 をまとめると　$\varepsilon_i \sim NID(0,\sigma^2)$

(X1) X_iと ε の無相関（X の外生性）:$Cov(X_i\varepsilon_i)=0$, $E(X_i\varepsilon_i)=X_iE(\varepsilon_i)=0$

2. 最小２乗推定量 a,b の性質

最小２乗推定量は **BLUE(Best Linear Unbiased Estimator,最良線形不偏推定量)**である

2.1 不偏性(unbiasedness)　p-28, 275

$E(b)=\beta+\Sigma w_iE(\varepsilon_i)=\beta$、$E(a)=\alpha$　← (U1),(X1)

2.2 有効性[1](最小分散性､効率性:efficiency)　p-31, 275

$$\sigma_b^2=E(b-E(b))^2=\sum w_t^2E(\varepsilon_t^2)=\sigma^2\sum w_t^2=\sigma^2\Big/\left(\sum(X_t-\bar{X})^2\right)$$ ← (U1),(U2),(U3),(X1)

$$\sigma_a^2=E(a-E(a))^2=E\left[(b-\beta)^2\bar{X}^2-(b-\beta)\bar{X}\bar{\varepsilon}+\bar{\varepsilon}^2\right]=\sigma^2\left[\frac{\bar{X}^2}{\sum(X_t-\bar{X})^2}+\frac{1}{n}\right]=\frac{\sigma^2\sum X_t^2}{n\sum(X_t-\bar{X})^2}$$

・任意の線形推定量 $C=c_0+\Sigma c_iY_i$ が不偏性 $E(C)=c_0+\Sigma c_i(\alpha+\beta X_i)=\beta$ を満たすには、$c_0=0$, $\Sigma c_i=0$, $\Sigma c_iX_i=1$ である必要がある →任意の線形不偏推定量は

$C=\Sigma c_iY_i$,　$\Sigma c_i=0$,　$\Sigma c_iX_i=\Sigma c_ix_i=1$

・C の分散 $var(C)=\sigma^2\Sigma c_i^2$ を最小にする c_i は　$L=\Sigma c_i^2-\lambda_1\Sigma c_i-\lambda_2(\Sigma c_ix_i-1)$ より $\partial L/\partial c_i=2c_i-\lambda_1-\lambda_2x_i=0$,　$\Sigma c_i=0$,　$\Sigma c_ix_i=1$ より求まる｡最初の条件 $c_i=(\lambda_1+\lambda_2x_i)/2$ を第２、３条件に代入すると $n\lambda_1+\lambda_2\Sigma x_i=0\rightarrow\lambda_1=0$、$\lambda_1\Sigma x_i+\lambda_2\Sigma x_i^2=2\rightarrow\lambda_2=2/(\Sigma x_i^2)$となり、このとき $c_i=x_i/\Sigma x_i^2=w_i$

2.3 Gauss-Markov Theorem　p-31

古典的回帰モデル((U-1),(U-2),(U-3),(X-1)が成り立つ)では最小２乗推定量は BLUE となる

2.4　$\varepsilon \sim NID(0,\sigma^2)$のときの最小２乗推定量 a、b の分布　p-30

$a\cdots N(\alpha,\sigma_a^2), b\cdots N(\beta,\sigma_b^2)$ だから、$Z=(b-\beta)/\sigma_b\cdots N(0,1)$　$\sigma_b=\sigma\Big/\sqrt{\sum(X_t-\bar{X})^2}$

このときσ²の不偏推定量 $s^2=\sum e_t^2/(n-k)$ で代用した $s_b=s\Big/\sqrt{\sum(X_t-\bar{X})^2}$ で標準化すると

[1] 不偏推定量の中で最も分散が小さい推定量を有効推定量(efficient estimator)という

$t=(b-\beta)/s_b$ は自由度 n-k の t 分布に従う（k はパラメタの数）。

3. 仮説検定　p-33

3.1 両側検定：リカードの等価定理の検証

①仮説の設定：$C_t=\beta_0+\beta_1Y_t+\beta_2B_t+\varepsilon_t$ において H_0: $\beta_2=0$　H1: $\beta_2\neq 0$

②帰無仮説の受容域：H_0 が正しいとしたら $b_2\sim N(0,\sigma_b)$だが $t=(b_2-0)/s_b=b_2/s_b\sim t(n-3)$となるので 1 - α の確率で統計量 t は $\Pr(-t_{\alpha/2}<t=\frac{b-\beta_0}{s_b}<t_{\alpha/2})=1-\alpha$ の範囲内に入る。

③仮説検定の実施：計算された t がこの受容域に入っている（いない）とき、**帰無仮説は有意水準 α で採択（棄却）される**、という。

※帰無仮説 H_0: $\beta=0$ の検定統計量 $t=b/s_b$ を「t 値 (t-ratio)」とよぶ。t 値が絶対値で 2 よりも大きいとき当該変数は統計的に有意だと考える（2 t ルール）。

3.2 片側検定：パンは必需財か否かの検定

①仮説の設定：$\ln Q_t=\beta_0+\beta_1\ln P_t+\beta_2\ln M_t+\beta_3Z_t+\varepsilon_t$ において H_0: $\beta_2=1$　H_1: $\beta_2<1$

②帰無仮説の受容域：H_0 が正しいとしたら $b_2\sim N(0,\sigma_b)$だが $t=(b_2-1)/s_b\sim t(n-4)$となるので 1 - α の確率で統計量 t は $\Pr(-t_{\alpha}<t=\frac{b-1}{s_b})=1-\alpha$ の範囲内に入る。

③仮説検定の実施：計算された t がこの採択域に入っている（いない）とき、**帰無仮説は有意水準 α で採択（棄却）される**。

　$t=(0.612-1)/0.149=-2.607>t_{5\%}(23-4)=-1.729$ より帰無仮説が棄却されるので、「パンは必需財である」といえる。

Excel の分析ツール(回帰分析)の結果（1980-2002 年）

回帰統計			係数	標準誤差	t	P-値	下限 95%	上限 95%
重相関 R	0.926	切片	3.970	1.552	2.558	0.019	0.722	7.218
重決定 R2	0.858	X 値 1	−1.009	0.135	−7.472	0.000	−1.292	−0.727
補正 R2	0.836	X 値 2	0.612	0.149	4.120	0.001	0.301	0.924
標準誤差	0.016	X 値 3	−0.461	0.310	−1.487	0.153	−1.109	0.188
観測数	23							

重決定 R2：決定係数 R^2, 補正 R2：自由度修正済み決定係数 $\overline{R}^2$、重相関 R：$\sqrt{R^2}$

標準誤差：$s=\sqrt{\sum e_i^2/(n-k)}$,p 値：H_0: $\beta=0$ のもとで t が生じる確率 $1-\Pr(-t<b/s_b<t)$

β の区間推定：$\Pr(b-t_{5\%/2}\cdot s_b<\beta<b+t_{5\%/2}\cdot s_b)=0.95$

4. 予測　p-38

4.1 条件付予測

点予測値 $\hat{Y}_t=a+bX_t$, 予測誤差 $e_t=Y_t-\hat{Y}_t=(\alpha-a)+(\beta-b)X_t+\varepsilon_t$

4.2 最良線形不偏予測量(Best Linear Unbiased Predictor)

$E(e_t)=0$, $E(\hat{Y}_t)=E(a)+E(b)X_t=\alpha+\beta X_t=E(Y_t)$ 不偏予測量

$Var(e_t)=E\{(\alpha-a)+(\beta-b)X_t\}^2+E(\varepsilon_t^2)=E(a-\alpha)^2+X_t^2E(b-\beta)^2+2X_tE\{(a-\alpha)(b-\beta)\}+E(\varepsilon_t^2)$

$=\sigma^2[1/n+(X_t-\overline{X})^2/\{\Sigma(X_i-\overline{X})^2\}=\sigma^2/n+\sigma^2_b(X_t-\overline{X})^2$

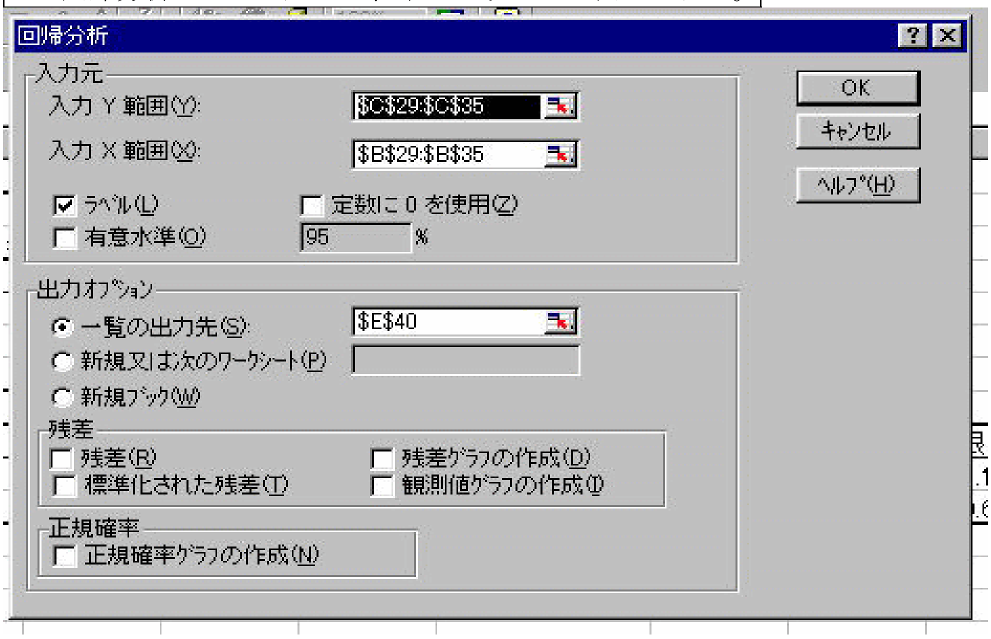
$Var(\hat{Y}_t)=\sigma^2[1+1/n+(X_t-\overline{X})^2/\{\Sigma(X_i-\overline{X})^2\}=\sigma^2(1+1/n)+\sigma^2_b(X_t-\overline{X})^2$

補足１　エクセルで回帰分析を行うには

１　ツール→分析ツール→回帰分析を選択します。

分析ツールが表示されない場合はツール→アドインで分析ツールをチェックしましょう

２　回帰分析ツールに入ると、下のような窓が表示される。



入力元

①入力 Y 範囲を指定：マウスで範囲指定します

②入力 X 範囲を指定：マウスで範囲指定します。複数の説明変数なら B29:D35 のように

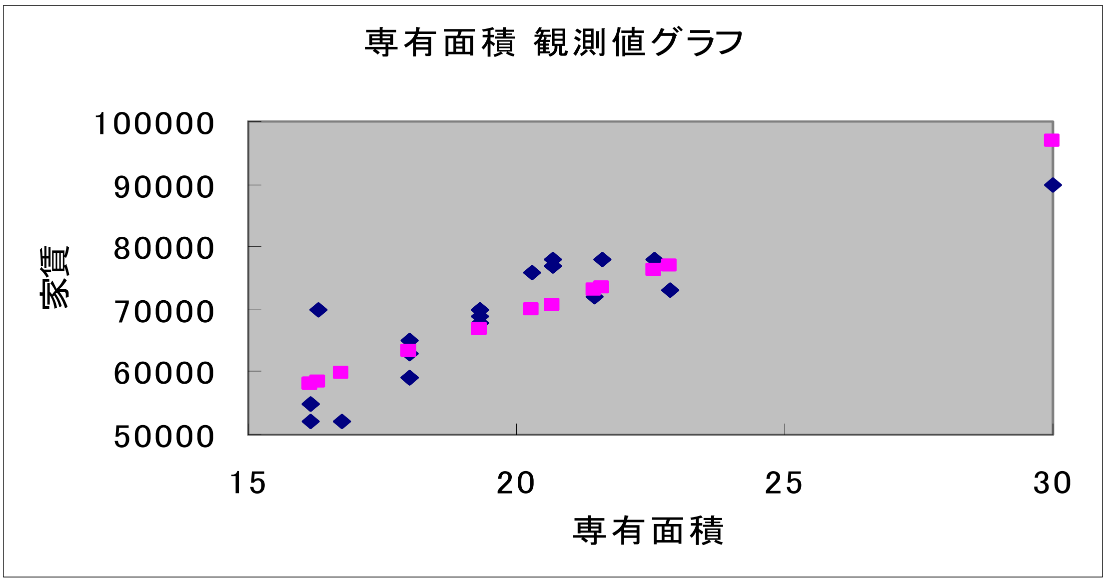
③変数名を入力範囲に加えてラベルにチェックをいれます。

出力オプション

①出力先：「新規又は次のワークシート」にしましょう。

②残差(R)、残差グラフ、観測値グラフにチェックを入れます。

３　推定結果の読み取り



概要		分散分析表					
回帰統計			自由度	変動	分散	分散比	有意 F
重相関 R	0.867103	回帰	1	1.7E+09	1.7E+09	63.6321	8.62E-08
重決定 R2	0.751867	残差	21	5.61E+08	26701760		
補正 R2	0.740051	合計	22	2.26E+09			
標準誤差	5167.375						
観測数	23						

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	12773.03	6893.564	1.852892	0.078004	-1562.92	27108.98
専有面積	2807.363	351.9333	7.976973	8.62E-08	2075.477	3539.248

(1)回帰統計（モデルの説明力をチェック）

重決定 R2：決定係数 R^2, 補正 R2：自由度修正済み決定係数 $\overline{R}^2$　($R^2 > \overline{R}^2$)

重相関 R：$\sqrt{R^2}$

標準誤差：s=$\sqrt{\sum e_i^2/(n-k)}$　残差 ε_i の母標準偏差 σ の不偏推定量

(2)分散分析表（データ変動の要因分解）テキスト p-79

TSS(合計変動, $\sum(Y_t-\overline{Y})^2$)=ESS(回帰変動, $\sum(\hat{Y}_t-\overline{Y})^2$)+RSS(残差変動, $\sum e_t^2$)

分散比＝{ESS/(k-1)}/{RSS/(n-k)}→H_0: $\beta_1=\beta_2=0$ の検定

有意 F：分散比が 0 となる確率

(3)回帰係数表（説明変数の影響力のチェック）

係数：β の最小２乗推定量 b

回帰係数の標準誤差：s_b=$\sqrt{s^2/\sum(X_i-\overline{X})^2}$

t 値：$t=\dfrac{b}{s_b}$　絶対値が２より大きいと有意（正確には t 分布表あるいは p 値を見る）

p 値：Ho: β=0 のもとで t が生じる確率　1－pr(-t<b/s_b<t)

下限/上限 95%：β の区間推定、Pr(b－$t_{2.5\%}$/s_b < β < b＋$t_{2.5\%}\cdot s_b$)=0.95